Nelson, R. M., Wallberg, A., Simões, Z. L. P., Lawson, D. J., & Webster, M. T. (2017). Genome-wide analysis of admixture and adaptation in the Africanized honeybee. *Molecular Ecology*, *26*(14), 3603-3617. https://doi.org/10.1111/mec.14122

Genome-wide analysis of admixture and adaptation in the Africanized honeybee

Ronald M. Nelson[1], Andreas Wallberg[1], Zilá Luz Paulino Simões[2], Daniel J. Lawson[3], Matthew T. Webster[1*]

1. Department of Medical Biochemistry and Microbiology, Science for Life Laboratory, Uppsala University, Uppsala, Sweden.

2. Department of Biology, University of São Paulo, São Paulo, Brazil.

3. Department of Mathematics, University of Bristol, Bristol, United Kingdom.

Keywords: Africanized honeybee, admixture, introgression, adaptation, biological invasion, natural selection

*matthew.webster@imbim.uu.se

Running title: Adaptation in Africanized honeybees

Abstract

Genetic exchange by hybridization or admixture can make an important contribution to evolution, and introgression of favourable alleles can facilitate adaptation to new environments. A small number of honeybees (*Apis mellifera*) with African ancestry were introduced to Brazil ~60 years ago, which dispersed and hybridized with existing managed populations of European origin, quickly spreading across much of the Americas. Here we analyse whole genome sequences of 32 Africanized honeybees sampled from throughout Brazil in order to study the effect of this process on genome diversity. By comparison with ancestral populations from Europe and Africa, we infer that these samples have 84% African ancestry, with the remainder from western European populations. However, this proportion varies across the genome and we identify signals of positive selection in regions with high European ancestry proportions. These observations are largely driven by one large gene-rich 1.4-Mbp segment on chromosome 11 where European haplotypes are present at a significantly elevated frequency and likely confer an adaptive advantage in the Africanized honeybee population. This region has previously been implicated in reproductive traits and foraging behaviour in worker bees. Finally, by analysing the distribution of ancestry tract lengths in the context of the known time of the admixture event, we are able to infer an average generation time of 2.0 years. Our analysis highlights the processes by which populations of mixed genetic ancestry form and adapt to new environments.

Introduction

Admixture and hybridisation can facilitate evolution as they lead to novel associations between genotypes that selection can act on (Anderson 1948; Anderson & Stebbins 1954; Stebbins 1959; Lewontin & Birch 1966; Abbott *et al.* 2013; Hedrick 2013). The mixture of different populations can also provide additional raw material for evolution by the transfer of adaptive alleles by gene flow. This process is potentially more efficient than adaptation from new mutations or standing variation. Alternatively, gene flow may have deleterious consequences. If genetic incompatibilities exist at certain loci then gene flow may be disadvantageous at these loci, leading to partial barriers to introgression (Barton & Hewitt 1985; Barton 2001). The mixture of populations is therefore now recognized as an important evolutionary process that can aid adaptation, dispersal and speciation (Stebbins 1959; Lewontin & Birch 1966; Hedrick 2013). It is now possible to analyse these processes on the fine scale by mapping local ancestry onto the genomes of admixed individuals and a number of statistical methods are available to do this (Liu *et al.* 2013).

In humans, admixture between modern and archaic humans appears to have provided a source of beneficial variants that led to improved survival in new environments (Racimo *et al.* 2015). For example, Neanderthals were a source of adaptive variation for skin phenotypes in modern humans (Sankararaman *et al.* 2014; Vernot & Akey 2014) and introgression with Denisovans facilitated adaptation to high altitudes in Tibetans (Huerta-Sánchez *et al.* 2014). Introgression has been shown to be an important source of adaptive alleles in many other species, including mimicry genes in *Heliconus*

butterflies (Zhang *et al.* 2016), genes involved in beak morphology in Darwin's finches (Grant & Grant 2002), yellow skin in domestic chickens (Hedrick 2013), melanism in gray wolves (Anderson *et al.* 2009) and environmental adaptation in maize (Hufford *et al.* 2013).

Africanized honeybees are an admixed population of the honeybee *Apis mellifera* that are now widespread in the Americas . The native distribution of *A. mellifera* encompasses Europe, Africa and the Middle East, and can be divided into five distinct clades, based on morphological and genetic data: A (African), C (Eastern Europe), M (Western Europe), O (Middle East), and Y (Ethiopia) (Ruttner 1988; Arias & Sheppard 1996; Franck *et al.* 2001; Harpur *et al.* 2014; Wallberg *et al.* 2014). This distribution likely reflects adaptation to local environmental conditions and hybridization between honeybee subspecies occurs where they meet along altitudinal and geographical clines (Sheppard *et al.* 1991; Oldroyd *et al.* 1995; Coroian *et al.* 2014). Honeybees in the Americas were initially composed of M and C clades introduced during and subsequent to the colonisation of the Americas by Europeans. The introduction of *A. mellifera* in South America, happened from the 1600s onwards, probably mainly involving honeybees from the M clade (Kerr 1967). Africanized honeybees originate from an introduction of African honeybees to Brazil in 1956 due to the accidental release of experimental populations into the wild in the vicinity of Sao Paulo. This source population was derived from 47 queens of the subspecies *A. m. scutellata*, 46 from Pretoria South Africa and one from Tanganyika (Kerr 1967; Scott Schneider *et al.* 2004).

The subsequent massive dispersal of Africanized honeybees in the Americas is a spectacular example of a rapid biological invasion. Although a proportion of their genetic ancestry is European, their phenotypic traits, which include their nesting biology, swarming and absconding behaviour, and foraging traits, are distinctly African (Roubik & Boreham 1990; Winston 1992). Due to the similar habitat and climate of their native range, Africanized honeybees are able to better survive in the wild than domestic European colonies. When Africanized honeybees came into contact with European colonies, hybridization and displacement has led to European traits being replaced by African ones (Sheppard *et al.* 1991; Rinderer *et al.* 1991; Clarke *et al.* 2002; Pinto *et al.* 2005; Rangel *et al.* 2016).

Africanized honeybees have now expanded dramatically to occupy an area from southern USA to northern Argentina and have largely replaced existing managed populations of European honeybees within this large geographical area. Their spread caused major problems for apiculture and had significant societal and economic impacts due to their highly aggressive stinging behaviour (Winston 1992; Scott Schneider *et al.* 2004; Ferreira *et al.* 2012; Byatt *et al.* 2015). However, the selective processes that occurred during the admixture and dispersal of Africanized honeybees and the reasons for their astounding success are not clear. Is there evidence that selection acted in favour of the spread of African alleles or against European ones during their dispersal? Why does some European ancestry persist in the population? Is there evidence for novel adaptations in the admixed population?

Here we analyse the genomes of 32 Africanized honeybees sampled throughout Brazil together with 98 genomes of African and European origin, sampled from their native

distribution (Wallberg *et al.* 2014). We use the method implemented in the program HAPMIX (Price *et al.* 2009) to infer patterns of local ancestry across the genome in the Africanized honeybees at high resolution and use this to perform several analyses of their population history and evolution. The genomic variation in ancestry proportions is informative about the selective pressures that occurred during the initial admixture and expansion and the demographic history. The results illuminate the processes that have shaped genome composition of Africanized honeybees and are relevant for our understanding of biological invasions.

Materials and methods

Sample collection and DNA extraction

New samples collected for this study comprise Africanized worker bees (n = 22) taken from unrelated colonies, during the last 3 years (Table S1, supporting information). The collection area spanned most of Brazil with several samples located close to Sao Paulo, near to where African honeybees were introduced (Figure S1, supporting information). We also included sequences from 10 further Africanized samples from a previous study in our analysis (see below).

We extracted DNA from the heads of individual worker bees using salt-ethanol precipitation. Brain tissue was dissolved in preparation buffer (100 mM NaCl, 10 mM Tris-HCl, pH = 8.0, 0.5% SDS) together with proteinase K at 50°C for 4 hours from individual heads cut in half. The samples were frozen overnight. DNA was precipitated

by adding saturated NaCl followed by 95% ethanol and spun into a pellet. The DNA pellet was suspended in TE buffer or double-distilled water and concentration fragment length assessed.

Sequencing and quality control

Samples were individually barcoded and whole genome sequencing was performed on the Illumina® HiSeq 2500 platform generating 2x125 bp paired-end reads to an average coverage of ~7.5× per sample (~170× in total across the 22 libraries). Reads were mapped to the Amel_4.5 assembly (Elsik *et al.* 2014) with BWA v0.7.12 (Li & Durbin 2010), followed by read-group and read-duplicate tagging with Picard v1.118 and indel-realignment and recalibration with GATK v3.3.0 (McKenna *et al.* 2010). We used the SNPs from Wallberg *et al*. (2014) to recalibrate quality scores in the new data.

Additional honeybee sequences

In addition to the new Africanized honeybee samples, we included genome sequences of 10 admixed Africanized individuals from a previous study (see Table S1, supporting information). We included 98 additional genomes from Wallberg *et al.* (2014), which represent four major honeybee subgroups and facilitate detection of admixed ancestry. These included individuals clustering within the M subgroup from *A. m. mellifera* (Norway; *n = 10*), *A. m. mellifera* (Sweden; *n = 9*) and *A. m. iberiensis* (Spain *n = 9*); within the C-subgroup from *A. m. carnica* (Austria *n = 10*) and *A. m. ligustica* (Italy *n = 10*); within the O-subgroup from *A. m. anatoliaca* (Turkey *n = 10*) and *A. m. syriaca* (Jordan *n*

= 10)*;* and from the A-subgroup from *A. m. adansonii* (Nigeria *n = 10*)*, A. m. scutellata* (South Africa *n = 10*)*, A. m. capensis* (South Africa *n = 10*). These libraries were sequenced using SOLiD™ technology. Short-read mapping and quality control were reused from the Wallberg *et al*. (2014) study for this part of the dataset.

SNP calling and haplotype phasing

Single-nucleotide polymorphisms (SNPs) were called across all 130 samples, comprising 22 Brazilian honeybee samples sequenced on Illumina for this study, 10 Brazilian honeybee samples sequenced on SOLiD as part of Wallberg *et al*. (2014) and 98 samples of mixed origin sequenced on SOLiD as part of Wallberg *et al*. (2014). We used the Bayesian population-based variant detection tool Freebayes v0.9.20-16 (Garrison & Marth 2012) for SNP calling. As the study was designed to detect regions of specific ancestry in the Brazilian honeybees by use of genetic variation in the native African and European populations from Wallberg *et al*. (2014) the SNP dataset was reduced to variants that had previously passed quality-filters in that study. We used BEAGLE v3.3.2 (Browning & Browning 2007) to phase haplotypes and impute missing variants. The genetic distance between Brazilian samples sequenced on SOLiD compared with Brazilian sequences on Illumina was 2-3 % higher than levels of genetic variation within these subsets of data, indicating that sequencing technology had minimal impact on estimated levels of genetic variation.

Analysis of Admixture

We performed unsupervised clustering using the program ADMIXTURE v1.23 (Alexander *et al.* 2009), which estimates the probability of assignment of unrelated individuals to *K* different populations. We used the full set of 6,823,740 SNPs (including all samples, *n = 130*) in 100 iterations at each *K* (from *K* 2-10). We used default settings except for generating a pseudo random seed using the system clock. We used the program CLUMPP v1.1 (Jakobsson & Rosenberg 2007) to determine the robustness of the assignments of individuals to populations at each *K*. First, we used CLUMPP to identify the common modes (patterns of assignment) among the 100 replicate ADMIXTURE runs at each *K.* The *LargeKGreedy* algorithm and 1000 random permutations were used as parameters in the CLUMPP analysis. We selected pairs with a symmetric similarity coefficient larger than 0.9 as indicative of a single mode. CLUMPP was run a second time with the most frequently occurring mode, using only the runs that belong to this mode (*LargeKGreedy* algorithm and 1000 random permutations). From this second analysis a mean clustering across the replicates for each K was obtained. We visualized the major modes of these clustering results using the program DISTRUCT (Jakobsson & Rosenberg 2007). Based on the results of the ADMIXTURE analysis, we classified the samples into 32 Africanized samples with mixed ancestry (all the samples from Brazil), 28 individuals representing the European genetic component (all samples from the M group), 30 individuals representing the African genetic component (all samples from the A group) and 40 individuals as genetic contrast population (C and O groups).

Dating Admixture

We estimated the number of generations since admixture in the Africanized population by applying an admixture linkage disequilibrium method implemented in *ROLLOFF* (Moorjani *et al.* 2011; Patterson *et al.* 2012). We used all samples from the M group as the first reference population, representing the European component, and all samples of *A. m. scutellata* as the second reference population, representing the African component. The rate of linkage disequilibrium decay in the admixed population between pairs of markers, weighted by the allele frequencies, is calculated in *ROLLOFF.* The number of generations since the admixture event is estimated by fitting an exponential distribution, by least squares, to the correlation of decay and distance between markers. To estimate the standard error, a weighted block jack-knife analysis was performed using ROLLOFF. Here one chromosome is excluded during each run which allows the stability of the date estimation to be evaluated.

F_{ST} analyses

Pairwise estimates of differentiation between populations were calculated using the Weir and Cockerham (1984) F_{ST} estimates at each SNP. The pairwise contrasts between the following groups were calculated: Group O, Group C, Group M and Group A. In addition, the pairwise contrasts between the Africanized population and the inferred ancestral populations (*A. m. scutellata* and Group M see above). These two sets of analyses were used to generate phylogenetic trees using the neighbour joining algorithm implemented in the APE package, implemented in R (Paradis *et al.* 2004).

Estimating Ancestry

We estimated variability in local ancestry across the genome of the Africanized population using HAPMIX (Price *et al.* 2009). We used the *A. m. scutellata* genome sequences to represent the African parental population and all the samples from group M to represent the European parental population. The phased genotypes for the admixed population were used. The average ancestry of the reference populations was set as θ = 20% European to African and the number of generations since admixture λ = 50. We obtained the likelihood for each SNP in each diploid individual that is was: 1. homozygous for African ancestry; 2. homozygous for European ancestry; 3. heterozygous for African/European ancestry. We summarized genome wide ancestry by averaging the local African ancestry proportion of each SNP across all individuals. Regions of high African and high European ancestry were defined to be areas containing consecutive SNPs with African average ancestry in the 99% quantile (high African ancestry) and in the 1% quantile (high European ancestry) respectively. Lastly we regressed the average level of African ancestry per individual on the sampling latitude.

To ensure there was no change in the ancestry probability assignment between our original estimates of 50 generations and later the 30 generation since admixture obtained by ROLLOFF, we re-run the HAPMIX analysis on a subset of the data. We found that all the SNPs assigned to either African or European ancestry had a perfect reassignment between the two estimates of λ. HAPMIX is known to be very robust in error for the λ and θ parameters (Price *et al.* 2009).

Population Ancestry model

We developed a method to detect SNPs with unusual ancestry proportions compared to the rest of the genome. We assume that individuals have an admixture fraction p_i and for each SNP they sample their ancestry independently. This gives rise to a Poisson-Binomial model for the total amount of a given ancestry A,

$$p(k_A) = PoiBin(k; \{p\}) = \prod_{i \in A} p_i \prod_{i \notin A} (1 - p_i).$$

This gives the expected distribution of k_A. We account for the correlation between SNPs by estimating the "effective number independent regions" *n*. This is done with the function "effectiveSize" from the R package "coda" (Plummer *et al.* 2006), which estimates how much a given ancestry *A(x)* decays over genetic distance *d* via $E_x(\lim_{d \to 0}(A(x+d) - A(x))/d)$. P-values are calculated from the tail area of this distribution. We can substitute *n* into the normal distribution approximation to the binomial for each *k* to obtain error bounds (*n* = 1300 here).

Simulated data for Poisson-Binomial model

We simulated data using ms (Hudson 2002) using the command line:

ms -N 1000 140 1 -t 1000 -r 2000 2000000 -I 3 40 40 60 -en 0 3 100 -es 0.05 3 0.2 -ej 0.05 3 1 -ej 0.05 4 2 -ej 0.5 2 1

This creates a population with 20% admixture from "Europe" and 80% from "Africa" 30 generations ago, which each split 300 generations ago, with $\mu = 0.0375$ and $\rho = 0.075$. The Africanized population has a large (666,667) effective population size. The Africanization date was rescaled to 30 generations ago by scaling the simulated

parameters. Firstly, multiplying N by r; secondly, dividing μ and ρ by r, and finally dividing the last split date by r where $r = 200/30$. We simulated 100 such regions of 2 Mb using this method and repeated the analysis above. Although this admixture simulation does not reflect Bee history as the African ancestry was selected to reach its current high proportion, this is exactly the signal that we hope to detect. Low population size was present in the African founders, but the method is sensitive only to low population size post-admixture, for which we see little evidence.

False Discovery rate (FDR) for the count data

For the purposes of this section, all loci with values in excess of the Poisson-Binomial are considered "true positives". We define the total fraction of positives in the dataset Q by the point where the observed distribution and the model prediction curves cross; further, let n_Q be the area under the model curve from 0 to Q, and m_Q be the additional area under the data curve. The number of positives is hence $P = m_Q$ and the number of negatives is $N = 1 - m_Q$. For a quantile q, the number of false positives is $FP = n_q$ and the number of true positives $TP = m_q$. The True Positive Rate is then FP/N and the False Positive Rate is TP/P. We report the true positive rate and the sensitivity, i.e. the fraction of positives that are true. The chosen quantile of 1% for low European ancestry/low African ancestry is a good choice with very low false positive rate, high enough true positive rate, high sensitivity and is symmetric for both ancestries.

Regions with evidence for selection

Selection for an adaptive variant is expected to reduce haplotype variation in flanking regions. To address whether variants and regions of interest had increased evidence for selection, we queried the 32 Africanized genomes for SNPs where one of the two alleles was associated with unusually long haplotypes compared to the other. To accomplish this analysis, we calculated the Integrated Haplotype Score (iHS)(Voight *et al.* 2006) for each SNP across the genome using the program SELSCAN v1.0.5 (Szpiech & Hernandez 2014). The iHS statistic is large and positive when a derived allele is found at high frequency and associated with low haplotype diversity over an extended region, which is indicative of a selective sweep. The iHS statistic can be negative if it is the ancestral allele that is associated with low haplotype diversity. A pairwise alignment was first produced between the two *Apis mellifera* (v4.5; (Elsik *et al.* 2014)) and *Apis cerana* reference genomes (v1.0; (Park *et al.* 2015)) using the SATSUMA whole-genome synteny package with default settings in order to annotate ancestral (0; negative iHS) and derived variants (1; positive iHS) at every SNP. Synteny was inferred for ~80% of the *A. mellifera* genome. For SNPs occurring outside of syntenic regions, the common variant among the 98 samples representing the diversity across the native range of the species (above) was taken as ancestral. A 100 kbp-resolution recombination map previously produced from African honeybees (Wallberg *et al.* 2015) was used as a source of recombination rates for the analyses. iHS was calculated for every putative haplotype core SNP with minor allele frequency > 0.02 (≥2/64 alleles) located within 10 kbp from its closest neighbour. Using this protocol, iHS was computed for 3,423,488 out of 4,224,690 sites (81%) segregating within the Brazilian population.

We next partitioned the data into 10 kbp windows and estimated the proportion of SNPs in each window that were associated with unusually high iHS values (|iHS|>2) as a means to assess local haplotype homozygosity across the genome while reducing the influence of linked signals between clustered SNPs. Windows were binned according to average ancestry scores (intervals of 10%) and we estimated mean iHS signals and 95% CI by bootstrapping each bin (2,000 replicates). This allowed us to test for significant enrichment of haplotype homozygosity in regions of high ancestry. In addition, we contrasted windows belonging to the top 1% percentile for high African and European ancestry, respectively, against the genomic background. The bootstrap procedure was repeated to test for enrichment of putatively functional elements in regions with high levels for either ancestry. For this test, we estimated the proportion of coding sequence in each 10 kbp window and compared ancestry outlier regions against the rest of the genome.

The gene models provided by the honeybee Official Gene Set v3.2 (Elsik *et al.* 2014) were used for all functional and gene-centric analyses. Scaffold-level gene coordinates from OGSv3.2 were converted into chromosome-level coordinates taking into account scaffold orientation listed in the Amel_v4.5 assembly using custom scripts.

Gene Ontology analysis

We performed a Gene Ontology (GO) analysis using the GOrilla platform (Eden *et al.* 2007, 2009). Genes within the high African ancestry and high European regions were identified separately. The *Drosophila* homologs of these genes were identified and used

as input gene lists in the GOrilla analysis tool. The *Drosophila* homologs of all matched honeybee genes were used as background gene list. Summaries of the GO terms were produced. The analysis was repeated for the genes in high European ancestry regions excluding chromosome 11.

Results

Admixed origin of Africanized honeybees

Here we analyse whole genome sequences of 32 Africanized honeybees, collected from 12 localities across Brazil (Figure S1, supporting information) together with a set of 100 samples taken from 10 different populations of honeybees from the A, C, M and O clades (Wallberg *et al.* 2014). The SNP dataset was pruned to include only variants that had previously been detected and passed quality filters in the Wallberg *et al*. (2014) study, which required SNP positions to have a minimum of 140x coverage. Missing genotypes were imputed in the entire dataset using Beagle. The resulting dataset had 6,823,740 SNPs distributed across the 16 nuclear chromosomes (see Methods).

We first used unsupervised clustering by ADMIXTURE (Alexander *et al.* 2009) to characterise ancestry proportions in Africanized honeybees relative to potential source populations in Africa and Europe (Figure 1). We estimated the robustness of the assignment per individual to a specific group using the program CLUMPP (Jakobsson & Rosenberg 2007). The proportion of replicates that supports the most common grouping of samples is referred to as the Assignment to the Major Mode (AMM). The

most robust grouping is at K = 3 (AMM = 100%, see Figure 1). At a value of K=4 (AMM = 70%), all of the major clades are identified and the Africanized sample contains 84.2% African ancestry, with most of the remainder, 15.0%, from the western European M clade (Figure 1, Table S2, supporting information). There is very little evidence for a genetic component from group C, 0.7%, or group O, <0.01% in the Africanized samples (Figure 1, Table S2, supporting information). Our analyses are therefore in concordance with previous reports indicating that Africanized honeybees in Brazil are composed of mainly group A descendants with some contribution from group M descendants, which possibly represents the main ancestry component of honeybees in Brazil prior to introduction of African honeybees (Clarke *et al.* 2001; Scott Schneider *et al.* 2004; Whitfield *et al.* 2006; Wallberg *et al.* 2014; Chapman *et al.* 2015). It is known that the originally introduced African honeybees were predominantly from the *A. m. scutellata* subspecies (47 queens in total)(Kerr 1967; Scott Schneider *et al.* 2004), although these are genetically highly similar to other African subspecies, and not possible to distinguish in an ADMIXTURE analysis (Wallberg *et al.* 2014).

We used the genetic distance based on average pairwise fixation index F_{ST} (Weir and Cockerham 1984) between the groups identified at K=4 to reconstruct their relationships as an evolutionary tree using the neighbour joining method (left side of Figure 2). We also created a neighbour joining tree using F_{ST} between our population samples of group M (n=39), *A. m. scutellata* (n=10) and the Africanized samples (n=32) to visualise the relationship of the admixed population relative to the contemporary populations of the ancestral groups (right side of Figure 2; F_{ST} per SNP between these contrasts are also displayed in Figure S2, supporting information).

Ancestry mapping to genome segments

We next mapped the average local African and European ancestry onto the genomes of Africanized samples using the program HAPMIX (Price *et al.* 2009). The *A. m. scutellata* and all M group honeybees were used as the two source populations. *A. m. scutellata* is the known subspecies that was introduced during the initial release of African honeybees. The specific European ancestry component is not clear, but inferred to be from the M group. The average local ancestry for the whole Africanized population was calculated for each SNP (Figure 3). The average proportion of African ancestry across all samples and all SNPs is 0.84 (mean = 0.844, median = 0.843, max = 1, min = 0.46). This level of African ancestry is almost exactly the same as that inferred using ADMIXTURE (0.844 vs. 0.842, respectively) or when estimated from window-based ancestries (0.844 vs. 0.838; 10 kbp windows). These ancestry proportions show more variation between SNPs (*SD* = 0.061) than between chromosomes (*SD* = 0.012) or samples (*SD* = 0.032) (Figure S3, supporting information). There is a segment on chromosome 11 with unusually low African (high European) ancestry compared to the rest of the genome and against simulated data (Figure 3b-c; additional analyses below).

We estimated the distribution of block lengths of European ancestry per individual from the HAPMIX results. Using only the consecutive European ancestry per genotype provides a conservative estimate (i.e. the shortest consecutive sections since disruptions by mixed ancestry are included) of the block length distribution (*mean* = 139 kbp, *median* = 86 kbp, see also Figure S4, supporting information). The longest block of

consecutive European ancestry is found on chromosome 6 in a single individual (1.3 Mbp).

Geographical variation in ancestry proportions

To investigate if the variation in ancestry proportions between the individuals is linked to a geographical pattern we correlated the proportion of European ancestry with longitude and latitude, as well as distance from the admixture epicentre. We found that there is no correlation between geography and ancestry in our samples, apart from a slight elevation in European ancestry in the most southern latitudes (R^2 = 0.302, p = 0.001, Figure S5, supporting information). This is consistent with previous studies (Sheppard *et al.* 1991; Scott Schneider *et al.* 2004; Whitfield *et al.* 2006; Abrahamovich *et al.* 2007) and suggests that Africanized honeybees have reduced fitness in the temperate climates leading to lower levels of African ancestry.

Hybridization time estimates

We used the software ROLLOFF to examine the decrease of admixture linkage disequilibrium in pairs of SNPs in the admixed population assuming a single pulse model of admixture (Moorjani *et al.* 2011). We examined the Africanized population with the European and African representative samples as estimates of the ancestral parental populations. The number of generations since admixture was obtained by fitting an exponential distribution to the ROLLOFF runs for each chromosome. The time since admixture is estimated to be 30.02 generations (*SE* = 1.86, Figure S6, supporting

information). The African honeybees were introduced to Brazil in 1956 (Scott Schneider *et al.* 2004), 59 years before sample collection, which enables estimation of a generation time of 2.0 (*SE* = 0.12) years.

Identification of genomic blocks with unusual ancestry proportions

We next investigated whether particular regions of the genome have an excess of either African or European ancestry that could indicate that they have been under selection. We identified local ancestry for all individuals and identified contiguous regions with the most extreme (1% quantile) proportions of each respective ancestry. Compared to the model, there is an excess of regions with both European and African ancestry (Figure 3 and Methods) which is well captured by the 1% quantile (high European ancestry) and 99% quantile (high African ancestry) thresholds. These thresholds are good choices to find candidate regions selected due to ancestry: the false positive rate is low (over 100 times better than chance) and sensitivity is high. Of the positives reported, we expect 92% and 64% respectively to be true (Methods; Figure S7, supporting information). The regions we report are thus likely to contain adaptive ancestral genotypes retained in the current environment. The one caveat is that our null model assumes no consanguinity in the Africanized honeybees, post-admixture event. Whilst population structure is inevitable, our model is only sensitive to sharing of recombination events that change ancestry and these are empirically rare.

Levels of high African ancestry, in 126 distinct blocks, appear to be distributed roughly uniformly across the genome (annotated in Table S3, supporting information; see also Figure 3 and Figure S2, supporting information). The size of the high African ancestry

peaks spans, on average, 10,380 bp (*median* = 5,876 bp, *SD* = 11,520 bp). In these regions, African alleles are fixed, or close to fixation, in the population. We find 103 distinct regions of high European ancestry across the genome. However, the population-wide level of European ancestry never exceeds 54% in any region on the genome and there are thus no fixed European derived alleles in this Africanized population (Figure 3, see also Figure S2, supporting information).

The average size of the high European ancestry peaks is 34,060 bp but the variation between the peak sizes is large (*median* = 7,924 bp, *SD* = 115,847 bp). The largest peak is found on chromosome 11 (from position 12.3 Mbp to 13.7 Mbp), spanning 1.4 Mbp (Figure 3). The magnitude of this peak is significantly beyond that expected by chance ($p = 5.6 \times 10^{-11}$; see Methods). This peak overlaps a region associated with ovary size and age of first foraging (see below). It is also interesting to note that there is a large block of consecutive and stable African ancestry on chromosome 11 (approximately 2 Mbp at coordinates 3.3 Mbp to 5.3 Mbp, Figure 4). This region has a low recombination rate relative to the rest of the genome estimated using both LD-based (Wallberg *et al.* 2015) and linkage-based (Ross *et al.* 2015) maps. In addition, we observe elevated levels of F_{ST}, especially from 4.1 - 5.3 Mb, in comparisons between non-admixed honeybee populations, including the parental populations (Figure 4), which suggests it is a region with high differentiation between these populations. We find 44 genes in the high African ancestry segments (30 have *Drosophila* orthologs, Table S4, supporting information). The high European ancestry regions contain a total of 378 genes (279 *Drosophila* orthologs) in the whole genome, 185 (134 *Drosophila* orthologs) of which are located on chromosome 11 (Table S3, supporting information).

A possible problem with our analysis is that regions of unusual ancestry might reflect regions where there is high differentiation between the ancestral populations. This could potentially generate a bias in ancestry proportions if assignment of ancestry is more accurate in these regions. In order to test for this, we estimated the F_{ST} per SNP between the inferred ancestral populations (samples of *A.m.scutellata* and the M group) as well as between the ancestral populations and the C group (visualised for the whole genome in Figure S2, supporting information). A correlation of the F_{ST} between the ancestral populations and the level of ancestry show that there is a very slight increase (*slope* = 0.07) in F_{ST} in regions with high African ancestry ($p = 2.2 \times 10^{-16}$, $R^2 = 0.008$, Figure S8, supporting information). Across the genome there also seems to be no indication that the either high European or high African ancestry peaks have high F_{ST} values co-located with it. An important implication is that our ability of detecting regions of biased ancestry does not seem to be dependent on having variants with different frequencies between populations and that regions of high F_{ST} between ancestral populations are not more likely to be identified as regions of biased ancestry.

Characterisation of a block of high European ancestry on chromosome 11

The most striking pattern in the data is a 1.4 Mbp block of high European ancestry on chromosome 11. The precise ancestry per individual varies across this region: three individuals have European ancestry blocks larger than 1 Mbp in this region, the largest being 1.2 Mbp, at coordinates 12.3 - 13.7 Mb (Figure 4 and Figure S4, supporting information). This 1.4 Mbp block comprises the majority of sequence of high European

ancestry in the genome (Figure 5) and contains 185 annotated genes. On a gene-by-gene basis, European ancestry ranges from 0.538 (GB45208) to 0.294 (GB45291) in this block. It contains 1.4% (n = 185) of all genes in the genome (n = 13,016; all genes across chromosomes 1-16 with ancestry scores) but 54% of the genes that occur within regions of high European ancestry (38.5x enrichment; Table S4, supporting information). Alleles and genes with European ancestry may therefore have been selectively favoured in this region.

Strikingly, the block on chromosome 11 overlaps a region that has been associated with reproductive and behavioural traits in honeybees in multiple studies. Linksvayer *et al* 2009 used backcrosses of Africanized and European honeybees to identify genetic markers associated with ovariole number in worker bees. We identified the location of the markers used in this study on the Amel_4.5 genome build by BLASTing their flanking sequences. The most significantly associated QTL identified by this study encompasses the block of European ancestry. Furthermore, another analysis of QTL mapping data, using reciprocal backcrosses of strains selected for high and low pollen hoarding (Rueppell 2009) identified an association with age of first foraging overlapping this block. The most strongly associated marker identified in this study (K118) is located at 13.0 Mbp on chromosome 11, which is in the centre of the European ancestry region in the vicinity of the mechanistic target of rapamycin (mTOR) gene, a key signalling and nutrient-sensing gene that controls larval growth and female caste differentiation in honeybees (Mutti *et al.* 2011). This region was also the most strongly associated QTL for ovary size in a similar backcross of pollen hoarding strains by (Ihle *et al.* 2015), marked in Figure 5.

The 1.4 Mbp block is 2.03x and 2.52x enriched for genes and coding sequence, respectively, compared to the rest of the genome (132 vs. 65 accessions per Mbp; 21.4% coding sequence vs. 8.5%; significantly higher than expected from randomly sampling 2,000 equally sized blocks; $p<0.05$; Figure 5). Among these, the proportion of genes with low CpG ratios in their coding sequences is unusually large ($CpG_{O/E}<1$; 62% vs. 45% in the rest of the genome; 1.37x enrichment; $p<0.05$; 2,000 random gene samples). $CpG_{O/E}$ across all coding sequence in this region, accordingly, is 0.84 compared to 1.04 across all coding sequence in the genome. Low CpG ratios indicate that the coding sequences in the region are likely subject to excess germline DNA methylation and reduced rates of recombination rates (Elango *et al.* 2009; Wallberg *et al.* 2015). Using the 100 kbp resolution population recombination map from African honeybees produced by Wallberg *et al.* (2015), we estimate the average recombination rate across the region to be reduced by 50% compared to the rest of the genome (13.45 cM/Mbp vs. 26.02 cM/Mbp; Figure 5).

The expected average ancestry block length can be estimated using the formula $L = [(1-m)r(t-1)]^{-1}$, where r is the recombination rate (crossovers per base pair per generation), m is the admixture proportion, and t is time since the admixture event in generations. (Racimo *et al.* 2015). Using estimates of $m = 0.16$ from the HAPMIX and ADMIXTURE analyses, $r = 2.2x10^{-7}$ from previous estimates on recombination rates (Beye *et al.* 2006; Wallberg *et al.* 2015; Liu *et al.* 2015) and $t = 30.02$ from our ROLLOFF estimate, gives an estimate of the average European haplotype length in Africanized honeybee population as 185,317 bp, which is slightly higher than the observed mean of 139 kbp (see above,

Figure S4, supporting information). We are interested in estimating the chance of finding the large European ancestry block on chromosome 11 as seen in this population. The chances finding a block of length a is calculated as: $\exp(-a/L)$ and therefore the probability of a block of 1.4 Mbp remaining since the admixture event is estimated as 0.0005. Given that the honeybee genome assembly is approximately 230 Mbp we estimate that the chance of observing this block in a single sample, by chance is 0.082 (~8%). The observation that we find such a block in the same position in the majority of the 32 samples can therefore be considered highly unlikely by chance alone.

We evaluated the possibility that the large section of continuous European ancestry on chromosome 11 could represent one or more structural variants between the two ancestral populations that supress recombination in this region. Several features argue against this interpretation. The borders of the European ancestry blocks are not consistent among samples, suggesting that recombination occurs within this block, which indicates the absence of discrete units of segregation in the population (Figure 4). We also examined the ancestry block borders in each sample visually using IGV. There were no clear breaks in continuity of pileup of reads at borders that could be associated with breakpoints of a structural element. The exact positions where the ancestry levels change were different between individuals but the reads at each of these positions mapped normally to the reference genome. Additionally, when comparing the F_{ST} between the ancestral populations and an outgroup (samples with group C ancestry) there is no indication that the section with higher European ancestry has a higher population differentiation compared to the average across the genome that could indicate a pre-existing structural element (Figure 4). We therefore do not find any

evidence to suggest that this block of unusual ancestry is connected to a structural variant.

Indications of positive selection in genomic regions with unusual ancestry proportions

We calculated the iHS statistic (Voight *et al.* 2006) for each SNP in the genome in order to test if the regions with elevated ancestry levels showed signals of selection based on haplotype length. We find significant enrichment for high iHS in regions of increasingly high European ancestry (Figure 5a). The top 1% windows (ancestry > 0.3304; n=200) are 3.29x enriched for iHS outlier SNPs compared to the rest of the genome ($p<0.05$, bootstrap). Out of the 200 windows, 57% belong to the 1.4 Mbp block on chromosome 11, whereas the remaining windows are distributed across 8 chromosomes (1, 2, 4, 5, 7, 10, 11 and 12, respectively). To assess whether the association between haplotype structure and ancestry was due to extreme signals in the single block, we removed this region from the genome and re-estimated the statistic. Although the most extreme regions are removed and the difference in haplotype homozygosity is reduced between outlier regions and the remaining genome, we still detect elevated iHS values with increasing European ancestry (top 1% windows enriched by 1.86x; $p<0.05$, bootstrap; Figure 6a). These results indicate that haplotype signals consistent with positive selection are found throughout many regions of high European ancestry but are the strongest in the 1.4 Mbp block on chromosome 11.

We repeated the iHS analysis with respect to extreme African ancestry but did not detect general enrichment of outlier SNPs among the top 1% 10 kbp windows (ancestry >

0.965; n = 200), but rather a small depletion (0.87x; $p < 0.05$, bootstrap). This indicates a lack of increased haplotype homozygosity in these regions and could indicate lack of positive selection on African alleles. However, the two 1% fractions represent very different deviation from the genome-wide average of either ancestry: the top 1% windows for European ancestry have levels 2.03x higher than the genomic background (mean = 0.329 vs. 0.162), whereas the corresponding top African regions are only elevated by 1.16x. It is therefore clear that the high average level of African ancestry obscures patterns of increased fixation and makes it difficult to localize putative targets of selection.

We next assessed evidence for functional enrichment in the high ancestry regions, taking the proportion of coding sequence in each 10 kbp window as a proxy for function and analysing the data as above. Under neutrality, we do not expect the proportion of genic elements to be related to ancestry. We detect a strong association between the proportion of coding sequence and European ancestry (Figure 6b) and find that the top 1% windows are 2.32x enriched for coding sequence ($p < 0.05$, bootstrap). After removing the gene-dense block on chromosome 11, we still observe a significant association ($p < 0.05$), although the enrichment in the most extreme regions is reduced to 1.45x. The regions with the highest levels of African ancestry are depleted of coding sequence compared to the rest of the genome (0.296x; $p < 0.05$, bootstrap).

Taken together, the observations that regions with high European ancestry tend to be enriched for functional sequence, and also enriched for SNPs with high iHS scores, indicates a tendency for functional alleles of European origin to be selectively favoured

in the Africanized population. However, these observations are predominantly driven by the extended block of high European ancestry on chromosome 11.

Gene ontology analysis

In order to gain further insight into the gene functions associated with either high African or European ancestry, we performed a gene ontology (GO) analysis of the genes within the outlier regions of either ancestry using the GOrilla platform (Eden *et al.* 2007, 2009). There were no significant enrichments detected for the regions of high African ancestry. We excluded chromosome 11 from the analysis of high European ancestry due to the presence of a large contiguous block of high European ancestry. In this analysis, we find significant enrichment for two GO terms related to olfaction: GO:0004984, olfactory receptor activity (*enrichment* = 11.22, FDR corrected *q-value* = 0.03) and GO:0005549, odorant binding (*enrichment* = 9.50, FDR corrected *q-value* = 0.04). In both cases, the GO term enrichment is due to the presence of the same 6 genes associated with odour binding (Table S5, supporting information). However, these genes are all closely located on chromosome 12 and this is therefore unlikely to represent independent enrichment of genes in a particular GO category.

Discussion

Here we have analysed whole genome sequences of 32 Africanized honeybees from Brazil compared with ancestral Old World populations in order to shed light on the processes involved in the invasion of Africanized honeybees. This allows us to study adaptation and migration in this admixed population on a fine scale by mapping ancestry onto the genome. First, by examining the fine-scale distribution of ancestry

blocks in the current Africanized honeybee population, we identify an excess of SNPs with high iHS values and high density of coding sequence in regions with high proportions of European ancestry. This indicates that some alleles at functional sites with European ancestry appear to be at a selective advantage in the Africanized population. This pattern is however mainly driven by a 1.4-Mbp block of European ancestry on chromosome 11 that segregates at high frequency in the population, likely due to positive selection. This region has been associated with reproductive and behavioural traits in multiple QTL studies. We also model the decay of linkage disequilibrium in the Africanized honeybee population to infer the number of generations since admixture. This allows us to estimate that the generation time for Africanized honeybees is about 2 years. These results allow insight into the ongoing evolutionary processes occurring in one of the most successful known biological invasions.

The role of natural selection in the Africanization process

A number of traits possessed by Africanized honeybees that are typical of honeybees from Africa are believed to have facilitated their successful spread. Africanized honeybees have faster colony growth and increased swarm production, partly because their colonies devote more space and resources to brood rearing (Roubik & Boreham 1990; Winston 1992). Africanized drones also have a mating advantage, mainly because they are produced in much greater numbers than European ones, leading to a greater transmission of African patrilines (Rinderer 1986). In conditions where there are multiple virgin queens in a honeybee colony, virgin queens with African paternity have

an advantage in the competition that ensues between them, partly due to their faster development (Winston 1992; Scott Schneider *et al.* 2004). Africanized honeybees sometimes also practice nest usurpation, a form of reproductive parasitism where they can replace existing colonies. However, Africanized honeybees are poorly adapted to survive cold winters, which European honeybees are able to survive by for example accumulating larger food stores and ceasing brood production, among other traits (Camazine & Morse 1988; Winston 1992). However, many traits that contribute to the success of Africanized honeybees may still be unknown.

The genomic distribution of ancestry proportions allows us to investigate the effect of natural selection in shaping genome variation. Against a background of predominantly African ancestry, we observe substantial variation in ancestry proportions across the genome. The Africanized population has low levels of European ancestry segregating across most of the genome. We identify signals that alleles of European ancestry may be favoured in the Africanized population. There is a slightly higher gene density and excess signals of positive selection in regions with high European ancestry proportions. This is consistent with the findings of (Zayed & Whitfield 2008), who found a higher proportion of European alleles among coding SNPs compared to noncoding SNPs in the Africanized population using a dataset of 444 SNPs. In our dataset, these patterns are mainly localized to a >1.4 Mbp block on chromosome 11 with elevated European ancestry. It is therefore likely that one or more variants of European ancestry in this region that are advantageous in the Africanized population.

The chromosome 11 block overlaps with QTLs identified in previous studies related to ovary size (Linksvayer *et al.* 2009; Ihle *et al.* 2015) and age of first foraging (Rueppell 2009; Linksvayer *et al.* 2009; Ihle *et al.* 2015). Diphenism in ovariole number between workers and queens is a major feature of social insects such as honeybees (Sherman *et al.* 1995). Furthermore, variation in ovariole number has been found by some studies to correlate with sensory tuning and foraging behaviour in worker bees: Workers with a lower number of ovarioles have been observed to begin foraging earlier, to be more likely to collect nectar rather that pollen, and to collect nectar with a higher sugar concentration compared to workers with more ovarioles (Amdam *et al.* 2004, 2006; Linksvayer *et al.* 2009). It has also been suggested that selection on reproductive behaviour in workers affects age of first foraging, but not foraging preferences (Oldroyd & Beekman 2008). A promising candidate gene found within this chromosome 11 region is mechanistic target of rapamycin (mTOR), which may influence ovary size during development (Linksvayer *et al.* 2009; Ihle *et al.* 2015). The TOR pathway also influences queen and worker developmental trajectories (Patel *et al.* 2007).

Genetic variants within this region may therefore affect both reproductive and foraging traits, and the abnormal levels of ancestry in this region in Africanized honeybees could indicate a selective advantage for European alleles controlling these traits. This suggests that Africanized honeybee biology is not simply equivalent to that of African honeybees and that some more typically European reproductive or foraging traits have a certain advantage in this population. On average, honeybee workers from African subspecies have a larger number of ovarioles, with the largest numbers exhibited by the Cape bee, *A. mellifera capensis* (Ruttner & Hesse 1981). In addition, honeybees with African

ancestry begin foraging earlier and tend to harvest more pollen (Winston *et al.* 1983). It is therefore possible that selection acted on this chromosome 11 region, which modulated these traits in the admixed Africanized population and facilitated adaptation to their new neotropical habitat.

Origins and timing of Africanization

The broad genetic composition of the Africanized bees corroborates the historical accounts of their origin and is consistent with previous analyses. Our analyses using both ADMIXTURE and HAPMIX (Alexander *et al.* 2009; Moorjani *et al.* 2011; Patterson *et al.* 2012) indicates that the genome of the Africanised population is composed of African and European admixture in an approximate 5:1 ratio. The African source population is difficult to distinguish from the analyses as African subspecies are very similar, but it is consistent with solely *A. m. scutellata* in concordance with historical data. As found in previous studies, we also observe that the European component is almost solely from the M group, from western Europe with little or no contribution from the C group from central Europe (Whitfield *et al.* 2006; Wallberg *et al.* 2014). The reason for the absence of C group alleles is unclear: it has been suggested that M group alleles may be favoured in the hybrid population, or that they are more compatible with African alleles (Smith *et al.* 1989; Clarke *et al.* 2001; Scott Schneider *et al.* 2004). However, it could however simply be the case that the European component of the original ancestral population of Africanized honeybees consisted solely of the M group and that bees with these ancestry proportions tend to replace managed populations of European origin.

After an initial admixture event, blocks of ancestry are broken up by recombination in each generation. The distribution of ancestry block lengths can therefore be used to infer the time since admixture. In most analyses this time is not known, but we know this accurately in the case of Africanized honeybees. We estimate the number of generations since the admixture event, which, given the time since African honeybee introduction allows us estimate the average generation time since this as 2.0 years (SE = 0.12). This is a useful parameter for population genetic inference, which is rarely possible to estimate in this way in any species. Whether this generation time is specific to Africanized honeybees or all honeybee populations is unclear. Africanized honeybees have been observed to swarm more often, approximately 5 times more that European honeybees (Benson 1985; Camazine & Morse 1988; Winston 1992), indicating that they invest more resources in reproduction, although it is unclear if this is related to generation time.

Conclusion

Here we show that the Africanized honeybee population in Brazil is predominantly composed of haplotypes of African origin. In general, the ancestry contributions from the European and African parental populations are relatively uniform across the genome, and analysing the distribution of ancestry blocks allows us to estimate the number of generations that have passed since the introduction of African honeybees to Brazil, which indicates a generation time of 2 years. We also identify a large 1.4 Mbp block of excess European ancestry on chromosome 11, which shows signs of being affected by natural selection. This region has been associated with the physiologically

connected traits of worker ovary size and foraging behaviour, which differ between European and African honeybees. It is therefore likely that adaptation in the Africanized honeybee population has involved these traits by natural selection acting on this genomic region. This study highlights how hybridization and admixture can facilitate adaptation.

Acknowledgements

This work was supported by the Swedish Research Council (2014-5096), the Swedish Research Council Formas (2013-722), the SciLifeLab Biodiversity Program (2014/R2-49) and the Carl Tryggers Stiftelse (CTS14:508).

References

Abbott R, Albach D, Ansell S *et al.* (2013) Hybridization and speciation. *Journal of Evolutionary Biology*, **26**, 229–246.

Abrahamovich AH, Atela O, Rúa PD la, Galián J (2007) Assessment of the mitochondrial origin of honey bees from Argentina. *Journal of Apicultural Research*, **46**, 191–194.

Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, **19**, 1655–1664.

Amdam GV, Csondes A, Fondrk MK, Page RE (2006) Complex social behaviour derived from maternal reproductive traits. *Nature*, **439**, 76–78.

Amdam GV, Norberg K, Fondrk MK, Page RE (2004) Reproductive ground plan may mediate colony-level selection effects on individual foraging behavior in honey bees. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 11350–11355.

Anderson E (1948) Hybridization of the habitat. Evolution, 2, 1–9.

Anderson E, Stebbins GL (1954) Hybridization as an evolutionary stimulus. Evolution, 8, 378.

Anderson TM, vonHoldt BM, Candille SI *et al.* (2009) Molecular and evolutionary history of melanism in North American gray wolves. *Science*, **323**, 1339–43.

Arias MC, Sheppard WS (1996) Molecular phylogenetics of honey bee subspecies (Apis mellifera L.) inferred from mitochondrial DNA sequence. *Molecular Phylogenetics and Evolution*, **5**, 557–66.

Barton NH (2001) The role of hybridization in evolution. *Molecular Ecology*, **10**, 551–568.

Barton NH, Hewitt GM (1985) Analysis of Hybrid Zones. *Annual Review of Ecology and Systematics*, **16**, 113–148.

Benson K (1985) The Africanized honey bee: genetic tactics of survival. The American Bee Journal, 125, 272-274.

Beye M, Gattermeier I, Hasselmann M *et al.* (2006) Exceptionally high levels of recombination across the honey bee genome. *Genome Research*, **16**, 1339–1344.

Browning SR, Browning BL (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics*, **81**, 1084–1097.

Byatt MA, Chapman NC, Latty T, Oldroyd BP (2015) The genetic consequences of the anthropogenic movement of social bees. *Insectes Sociaux*, **63**, 15–24.

Camazine S, Morse RA (1988) The Africanized honeybee. *American Scientist*, **76**, 464–471.

Chapman NC, Harpur BA, Lim J *et al.* (2015) A SNP test to identify Africanized honeybees via proportion of "African" ancestry. *Molecular Ecology Resources*, **15**, 1346–1355.

Clarke KE, Oldroyd BP, Javier J, Quezada-Euán G, Rinderer TE (2001) Origin of honeybees (Apis mellifera L.) from the Yucatan peninsula inferred from mitochondrial DNA analysis. *Molecular Ecology*, **10**, 1347–1355.

Clarke KE, Rinderer TE, Franck P, Quezada-Euán JG, Oldroyd BP (2002) The Africanization of honeybees (Apis mellifera L.) of the Yucatan: a study of a massive hybridization event across time. Evolution, 56, 1462–1474.

Coroian CO, Muñoz I, Schlüns EA *et al.* (2014) Climate rather than geography separates two European honeybee subspecies. *Molecular Ecology*, **23**, 2353–2361.

Eden E, Lipson D, Yogev S, Yakhini Z (2007) Discovering motifs in ranked lists of DNA sequences. PLoS Computational Biology, 3, e39.

Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z (2009) GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, **10**, 48.

Elango N, Hunt BG, Goodisman MA, Yi SV (2009) DNA methylation is widespread and associated with differential gene expression in castes of the honeybee, Apis mellifera. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 11206–11.

Elsik CG, Worley KC, Bennett AK et al. (2014) Finding the missing honey bee genes: lessons learned from a genome upgrade. BMC Genomics, 15, 86.

Ferreira RS, Almeida RAMB, Barraviera SRCS, Barraviera B (2012) Historical Perspective and Human Consequences of Africanized Bee Stings in the Americas. *Journal of Toxicology and Environmental Health, Part B*, **15**, 97–108.

Franck P, Garnery L, Loiseau A et al. (2001) Genetic diversity of the honeybee in Africa: microsatellite and mitochondrial data. Heredity, 86, 420–30.

Garrison E, Marth G (2012) Haplotype-based variant detection from short-read sequencing. *arXiv:1207.3907 [q-bio]*.

Grant PR, Grant BR (2002) Unpredictable Evolution in a 30-Year Study of Darwin's Finches. *Science*, **296**, 707–711.

Harpur BA, Kent CF, Molodtsova D et al. (2014) Population genomics of the honey bee reveals strong signatures of positive selection on worker traits. Proceedings of the National Academy of Sciences, 111, 2614-2619.

Hedrick PW (2013) Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. *Molecular Ecology*, **22**, 4606–4618.

Hudson RR (2002) Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics*, **18**, 337–338.

Huerta-Sánchez E, Jin X, Asan *et al.* (2014) Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature*, **512**, 194–197.

Hufford MB, Lubinksy P, Pyhäjärvi T *et al.* (2013) The Genomic Signature of Crop-Wild Introgression in Maize (R Mauricio, Ed,). *PLoS Genetics*, **9**, e1003477.

Ihle KE, Rueppell O, Huang ZY et al. (2015) Genetic Architecture of a Hormonal Response to Gene Knockdown in Honey Bees. Journal of Heredity, 106, 155-165

Jakobsson M, Rosenberg NA (2007) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*, **23**, 1801–1806.

Kerr WE (1967) The history of the introduction of African bees to Brazil. *South African Bee Journal*, **39**, 3–5.

Lewontin RC, Birch LC (1966) Hybridization as a Source of Variation for Adaptation to New Environments. *Evolution*, **20**, 315–336.

Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, **26**, 589–595.

Linksvayer TA, Rueppell O, Siegel A et al. (2009) The genetic basis of transgressive ovary size in honeybee workers. Genetics, 183, 693–707.

Liu Y, Nyunoya T, Leng S *et al.* (2013) Softwares and methods for estimating genetic ancestry in human populations. *Human Genomics*, **7**, 1.

Liu H, Zhang X, Huang J *et al.* (2015) Causes and consequences of crossing-over evidenced via a high-resolution recombinational landscape of the honey bee. *Genome Biology*, **16**, 1.

McKenna A, Hanna M, Banks E *et al.* (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**, 1297–1303.

Moorjani P, Patterson N, Hirschhorn JN et al. (2011) The History of African Gene Flow into Southern Europeans, Levantines, and Jews. PLoS Genetics, 7, e1001373.

Mutti NS, Dolezal AG, Wolschin F *et al.* (2011) IRS and TOR nutrient-signaling pathways act via juvenile hormone to influence honey bee caste fate. *Journal of Experimental Biology*, **214**, 3977–3984.

Oldroyd BP, Beekman M (2008) Effects of Selection for Honey Bee Worker Reproduction on Foraging Traits. *PLOS Biology*, **6**, e56.

Oldroyd BP, Cornuet J-M, Rowe D, Rinderer TE, Crozier RH (1995) Racial admixture of Apis mellifera in Tasmania, Australia: similarities and differences with natural hybrid zones in Europe. Heredity, 74, 315–325.

Paradis E, Claude J, Strimmer K (2004) APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*, **20**, 289–290.

Park D, Jung JW, Choi B-S *et al.* (2015) Uncovering the novel characteristics of Asian honey bee, Apis cerana, by whole genome sequencing. *BMC Genomics*, **16**, 1.

Patel A, Fondrk MK, Kaftanoglu O *et al.* (2007) The Making of a Queen: TOR Pathway Is a Key Player in Diphenic Caste Development. *PLOS ONE*, **2**, e509.

Patterson N, Moorjani P, Luo Y et al. (2012) Ancient admixture in human history. Genetics, 192, 1065–1093.

Pinto MA, Rubink WL, Patton JC, Coulson RN, Johnston JS (2005) Africanization in the United States: replacement of feral European honeybees (Apis mellifera L.) by an African hybrid swarm. *Genetics*, **170**, 1653–1665.

Plummer M, Best N, Cowles K, Vines K (2006) CODA: Convergence diagnosis and output analysis for MCMC. *R news*, **6**, 7–11.

Price AL, Tandon A, Patterson N et al. (2009) Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS Genetics, 5, e1000519.

Racimo F, Sankararaman S, Nielsen R, Huerta-Sánchez E (2015) Evidence for archaic adaptive introgression in humans. *Nature Reviews Genetics*, **16**, 359–371.

Rangel J, Giresi M, Pinto MA *et al.* (2016) Africanization of a feral honey bee (Apis mellifera) population in South Texas: does a decade make a difference? *Ecology and Evolution*, **6**, 2158–2169.

Rinderer TE (1986) Africanized bees: an overview. American Bee Journal, 126, 98-100; 128-129.

Rinderer TE, Stelzer JA, Oldroyd BP, Buco SM, Rubink WL (1991) Hybridization between European and Africanized honey bees in the neotropical Yucatan Peninsula. Science, 253, 309–311.

Ross CR, DeFelice DS, Hunt GJ *et al.* (2015) Genomic correlates of recombination rate and its variability across eight recombination maps in the western honey bee (Apis mellifera L.). *BMC Genomics*, **16**, 107.

Roubik D, Boreham M (1990) Learning to live with Africanized honeybees. *Interciencia*, **15**, 146–153.

Rueppell O (2009) Characterization of quantitative trait loci for the age of first foraging in honey bee workers. Behavior Genetics, 39, 541–553.

Ruttner F (1988) *Biogeography and Taxonomy of Honeybees*. Springer-Verlag, Berlin.

Ruttner F, Hesse B (1981) Rassenspezifische unterschiede in ovarentwicklung und eiablage von weisellosen arbeiterinnen der honigbiene Apis mellifera L. Apidologie, 12, 159–183.

Sankararaman S, Mallick S, Dannemann M *et al.* (2014) The genomic landscape of Neanderthal ancestry in present-day humans. *Nature*, **507**, 354–357.

Scott Schneider S, DeGrandi-Hoffman G, Smith DR (2004) The African honey bee: Factors Contributing to a Successful Biological Invasion. Annual Review of Entomology, 49, 351–376.

Sheppard WS, Rinderer TE, Mazzoli JA, Stelzer JA, Shimanuki H (1991) Gene flow between African- and European-derived honey bee populations in Argentina. *Nature*, **349**, 782–784.

Sherman PW, Lacey EA, Reeve HK, Keller L (1995) The eusociality continuum. Behavioral Ecology, 6, 102–108.

Smith DR, Taylor OR, Brown WM (1989) Neotropical Africanized honey bees have African mitochondrial DNA. *Nature*, **339**, 213–215.

Stebbins GL (1959) The role of hybridization in evolution. *Proceedings of the American Philosophical Society*, **103**, 231–251.

Szpiech ZA, Hernandez RD (2014) selscan: An efficient multithreaded program to perform EHH-based scans for positive selection. Molecular Biology and Evolution, 31, 2824-2827.

Vernot B, Akey JM (2014) Resurrecting surviving Neandertal lineages from modern human genomes. Science, 343, 1017–1021.

Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. PLOS Biology, 4, e72.

Wallberg A, Glémin S, Webster MT (2015) Extreme recombination frequencies shape genome variation and evolution in the honeybee, Apis mellifera. PLOS Genetics, 11, e1005189.

Wallberg A, Han F, Wellhagen G *et al.* (2014) A worldwide survey of genome sequence variation provides insight into the evolutionary history of the honeybee Apis mellifera. *Nature Genetics*, **46**, 1081–1088.

Whitfield CW, Behura SK, Berlocher SH *et al.* (2006) Thrice out of Africa: ancient and recent expansions of the honey bee, Apis mellifera. *Science*, **314**, 642–645.

Winston ML (1992) The biology and management of Africanized honey bees. Annual Review of Entomology, 37, 173–193.

Winston ML, Taylor OR, Otis GW (1983) Some differences between temperate European and tropical African and South American honeybees. *Bee World*, **64**, 12–21.

Zayed A, Whitfield CW (2008) A genome-wide signature of positive selection in ancient and recent invasive expansions of the honey bee Apis mellifera. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 3421–6.

Zhang W, Dasmahapatra KK, Mallet J, Moreira GRP, Kronforst MR (2016) Genome-wide introgression among distantly related Heliconius butterfly species. *Genome Biology*, **17**.

Data Accessibility

All data from this study are available at the NCBI Sequence Read Archive (SRA) under BioProject IDs PRJNA236426 and PRJNA350769.

Author Contributions

MTW designed the study. RMN, AW and DL performed the analysis. ZLPS contributed samples. MTW, RMN and AW wrote the paper with contributions from all authors.

Figure legends

Figure 1. Population structure of the 130 individuals based on ~6.8*M* SNPs and assuming 2 to 7 clusters using the program ADMIXTURE. Groups are shown above and subspecies names are shown below. Only the major modes are shown (AMM shown in brackets). All Old World populations are partitioned according the four major lineages (A, M, C, and O) at K=4 and the Africanized population appears to have 84% Africanized ancestry with the remainder from the M group. Africanized honeybees are assigned to their own distinct ancestry group at K=5 and greater. Key: *A. m. melifera* – mel (from Norway and Sweden respectively – NO; SE); *A. m. Iberiensis* - ibe; *A. m. adansonii* - ada; *A. m. scutelata* - scu; *A. m. capensis* - cap; *A. m. anatoliaca* - ana; *A. m. syrica* - syr; *A. m. carnica* - car; *A. m. ligustica* – lig.

Figure 2. Phylogenies based on the F_{ST} contrasts. Left side (phylogeny in black) – main honeybee group differentiation including the relative separation of the ancestral groups of the Africanized population. Right side (phylogeny in red and blue)– relative differentiation of the ancestral populations and the admixed Africanized population. Scale: Pairwise F_{ST}. Note that the African component of the Africanized honeybees is considered to come solely from the *A. m. scutellata* population in this tree. Key: *A. m. melifera* – mel; *A. m. Iberiensis* - ibe; *A. m. adansonii* - ada; *A. m. scutelata* - scu; *A. m. capensis* - cap; *A. m. anatoliaca* - ana; *A. m. syrica* - syr; *A. m. carnica* - car; *A. m. ligustica* – lig.

Figure 3. a) Average proportion of African ancestry of SNPs across the whole genome for the Africanized honeybee population inferred using the program HAPMIX. Grey horizontal lines indicate the top 99% quantile and bottom 1% quantile relating to high African ancestry and high European ancestry regions respectively. A prominent drop in African ancestry spanning a region of 1.4 Mb is observed on chromosome 11 b) A histogram of this data (black, with the exclusion of the significant region in Chr11 as dotted), compared to the Poisson-Binomial prediction (red line, error bars as dashed). Note that the histogram density is on a log-scale. c) As for b) for simulated data, showing that the model fits.

Figure 4. Plots of ancestry proportions and population differentiation across chromosome 11. a) Top panel – F_{ST} contrasts between ancestral populations (Group M, and *A. m. scutellata*) and an unrelated group (Group C). Bottom panel – Average proportion of African ancestry. Grey horizontal lines indicate the top 99% quantile and bottom 1% quantile relating to high African ancestry and high European ancestry regions respectively. Bar on x-axis indicates the marker density. b) Relative ancestry at each SNP for each individual (individual identifiers correspond to those in Table S1, supporting information). Red, probability of African ancestry at both alleles; Blue, probability of European ancestry at both alleles and Grey, probability of one African and one European allele. A 1.4 Mbp block of consecutive high European ancestry is marked within vertical black lines.

Figure 5. Proportion of European ancestry along chromosome 11 (10 kbp windows; blue barplot; left y-axis). A 1.4Mbp block of high European ancestry (coordinates 12,3-13,7Mbp) is highlighted in light blue. Gene positions and their respective European ancestry are shown as circles (mid-point coordinates; green=high CpG genes; yellow=low CpG genes). The genome-wide top 1% percentile for European ancestry is indicated with a white dashed line. Recombination rates are shown across the region (100 kbp windows; dark dotted line; right y-axis). Shaded top bar and region indicates approximate intervals for a QTL region for ovary size from (Ihle *et al.* 2015). Black triangles represents current positions of significant QTL markers K7714 (12,597,388), K17285 (12,880,427), K18146 (14,129,134), K3405 (14,000,253) and K16813 (14,447,359) from that study.

Figure 6. a) The proportion of SNPs with high haplotype homozygosity scores (|iHS|>2) in 10 kbp windows with different levels of European ancestry (light green=all data; dark green=data after removing the 1.4 Mbp block on chromosome 11; error bars indicate 95% confidence intervals estimated from 2,000 bootstrap replicates). b) The proportion of coding sequence in 10 kbp windows in relation to European ancestry (light blue=all data; dark blue=data after removing the chromosome 11 block; errors bars computed as in a).

Supporting Information

Figure S1. Sampling locations. Map of sampling locations of Africanized honeybees in Brazil. A total of 32 samples from 12 localities were included (AL - Alfenas; AP - Apiai;

LA - Luiz Antonio; MA - Manaus; M - Mossoro; PG - Pindamonhangaba; Q - Querencia; RG - Rio Grande do Sul; RI - Rio Claro; RP - Ribeirao Preto; X - Xanxerê; SP - Sao Paulo).

Figure S2. Chromosome wide plots of ancestry proportions and population differentiation. Genome wide analysis. Top – F_{ST} contrasts for the ancestral populations (Group M, and *A. m. scutellata*) compared to the F_{ST} of an unrelated group (Group C). Middle -Average proportion African ancestry for chromosome 11. Grey horizontal lines indicate the top 99% quantile and bottom 1% quantile relating to high African ancestry and high European ancestry regions respectively. Bar on x-axis indicated the marker density. Bottom – Relative ancestry at each SNP for each individual (not same scale as top). Red, probability of African ancestry at both alleles; blue probability of European ancestry at both alleles and grey, probability of one African and one European allele.

Figure S3. Average ancestry per chromosome inferred by HAPMIX across 32 samples. The distribution of the proportion African ancestry for each chromosome given the variation between individual samples. The whiskers show the 1.5 inter quartile range of the data.

Figure S4. European tract length distribution per chromosome. Green line - mean track length, Yellow line – median track length.

Figure S5. Average ancestry per sample plotted against sampling latitude. The dotted line indicates the latitude of Sao Paulo, which is close to the site of the initial release of African honeybees. The increase in African ancestry is significantly correlated with the increase in latitude (P=0.011) and is particularly reduced south of Sao Paulo.

Figure S6. Weighed correlation of LD between pairs of markers from parental populations as a function of genetic distance as calculated by ROLLOFF

Figure S7. a) Receiver-operator curves (ROC) showing the relationship between the fraction of positives that are found (True Positive Rate) and the number of negatives that are falsely found (False Positive Rate), for different choices of quantile. Shown is the expected fraction of true vs the expected number of false positives as calculated under the Poisson-Binomial model. The piecewise-linear curve arises due to counting discrete individuals in the data. The 1% quantile threshold in both the high European (black) and high African (red) case appears as a good choice in terms of favouring true positives. b) The sensitivity, or fraction of reported positives that are actually true, as a function of the false positive rate showing the chosen 1% quantile region. The perfect precision region for Europeans corresponds to the reported region on Chromosome 11.

Figure S8. Average F_{ST} per African ancestry level where the African ancestry is divided into 20 levels.

Supplementary Tables

Table S1. Sample description including group name, group membership, sampling location and genotyping platform used for sequencing.

Table S2. Admixture assignment of the Africanized individuals to the four major groups specified at K4. Values obtained from the Clump consensus across 100 admixture runs.

Table S3. Regions within the high (1% quantile) European and African ancestry respectively.

Table S4. Genes within the high (1% quantile) European and African ancestry respectively including *Apis mellifera* genes and *Drosophila melanogaster* orthologs within each region. Consecutive high European ancestry region highlighted in yellow on chromosome 11 give (from position 12.3 Mbp to 13.7 Mbp).

Table S5. Genes from high European ancestry regions (excluding chromosome 11), in enriched GO terms.